(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 737 927 B1

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
**05.12.2001 Bulletin 2001/49**

(51) Int Cl.[7]: **G06F 17/27**, G06F 17/30

(21) Application number: **96302250.4**

(22) Date of filing: **29.03.1996**

(54) **Automatic method of generating thematic summaries**

Automatisches Verfahren zur Erzeugung von thematischen Zusammenfassungen

Méthode automatique pour la génération de résumés thématiques

(84) Designated Contracting States:
**DE FR GB**

(30) Priority: **14.04.1995 US 422573**

(43) Date of publication of application:
**16.10.1996 Bulletin 1996/42**

(73) Proprietor: **XEROX CORPORATION**
**Rochester, New York 14644 (US)**

(72) Inventor: **Chen, Francine R.**
**Menlo Park, CA 94025 (US)**

(74) Representative: **Grünecker, Kinkeldey,**
**Stockmair & Schwanhäusser Anwaltssozietät**
**Maximilianstrasse 58**
**80538 München (DE)**

(56) References cited:
**US-A- 5 384 703**

- **EDMUNDSON, H.P.: "New methods in automatic extracting" JOURNAL OF THE ASSOCIATION FOR COMPUTING MACHINERY, vol. 16, April 1969, pages 264-285, XP002078269**
- **"METHOD FOR AUTOMATIC EXTRACTION OF RELEVANT SENTENCES FROM TEXTS" IBM TECHNICAL DISCLOSURE BULLETIN, vol. 33, no. 6A, November 1990, page 338/339 XP002015802**
- **BLACK W J ET AL: "A PRACTICAL EVALUATION OF TWO RULE-BASED AUTOMATIC ABSTRACTING TECHNIQUES" EXPERT SYSTEMS FOR INFORMATION MANAGEMENT, vol. 1, no. 3, 1988, pages 159-177, XP002015761**
- **LUHN, H.P.: "The automatic creation of literature abstracts" IBM JOURNAL,April 1958, page 159-165 XP002078270**



EP 0 737 927 B1

## Description

**[0001]** The present invention relates to a method of automatic text processing. In particular, the present invention relates to an automatic method of generating thematic summaries of documents.

**[0002]** Document summaries and abstracts serve a valuable function by reducing the time required to review documents. Summaries and abstracts can be generated after document creation either manually or automatically. Manual summaries and abstracts can be of high quality but may be expensive because of the human labor required. Alternatively, summaries and abstracts can be generated automatically. Automatic summaries and abstracts can be cheaper to produce, but obtaining high quality consistently is difficult

**[0003]** Systems for generating automatic summaries rely upon one of two computational techniques, natural language processing or quantitative content analysis. Natural language processing is computationally intensive. Additionally, producing semantically correct summaries and abstracts is difficult using natural language processing when document content is not limited.

**[0004]** Quantitative content analysis relies upon statistical properties of text to produce summaries. Gerald Salton discusses the use of quantitative content analysis to summarize documents in "Automatic Text Processing" (1989). The Salton summarizer first isolates text words within a corpus of documents. Next, the Salton summarizer flags as title words words used in titles, figures, captions, and footnotes. Afterward, the frequency of occurrence of the remaining text words within the document corpus is determined. The frequency of occurrence and the location of text words are then used to generate word weights. The Salton summarizer uses the word weights to score each sentence of each document in the document corpus. These sentence scores are used in turn to produce a summary of a predetermined length for each document in the document corpus. Summaries produced by the Salton summarizer may not accurately reflect the themes of individual documents because word weights are determined based upon their occurrence across the document corpus, rather than within each individual document.

**[0005]** Edmundson, H. P.: "New Methods In Automatic Extracting" Journal of the Association For Computing Machinery, vol. 16, April 1969, pages 264-285, XP002078269 discloses methods of automatically extracting documents for screening purposes, i.e. the computer selection of sentences having the greatest potential for conveying to the reader the substance of the document. The described methods treat in addition to the presence of high-frequency content words (keywords) three components: pragmatic words (cue words); title and heading words; and structural indicators (sentence location).

**[0006]** US-A-5 384 703 describes automatically forming a summary by selecting regions of a document. Each selected region includes at least two members of a seed list. The seed list is formed from a predetermined number of the most frequently occurring complex expressions in the document that are not on a stop list. If the summary is too long, the region selection process is performed on the summary to produce a shorter summary. This region-selection process is repeated until a summary is produced having a desired length. Each time the region selection process is repeated, the seed list members are added to the stop list and the complexity level used to identify frequently occurring expressions is reduced.

**[0007]** "Method for Automatic Extraction Of Relevant Sentences From Texts", IBM Technical Disclosure Bulletin, vol. 33, no 6A, November 1990, page 338-339, XP002015802 describes a method for automatically extracting from a text in any language the most significant and explicative sentences. All the words of the text are analyzed and listed according to an order of importance. Only useful words are considered excluding "stop words" such as articles, pronouns, prepositions, etc. The importance of selected words is evaluated according to their frequency and position within the text. The words with high-frequency and located in the title, introduction or conclusion are considered of increased importance.

**[0008]** Black W. J. et al.: "A Practical Evaluation Of Two Rule-Based Automatic Abstracting Techniques", Expert Systems For Information Management, vol. 1, no 3, 1988, pages 159-177, XP002015761 describes automatic abstracting techniques based on statistical analyzes and superficial syntactic pattern matching of texts, which lend themselves to implementation using a rule-based approach.

**[0009]** Luhn, H. P.: "The Automatic Creation Of Literature Abstracts", IBM Journal, April 1958, pages 159-165, XP002078270 describes a method for automatically creating an abstract from an article presented in machine-readable form to a data-processing machine, whereby statistical information is derived from word frequency and distribution. This statistical information is used by the machine to compute a relative measure of significance, first for individual words and then for sentences. Sentences scoring highest in significance are extracted and printed out to become the "auto-abstract".

**[0010]** An object of the present invention is to automatically generate improved document summaries that accurately reflect the themes of each document.

**[0011]** According to the present invention, there is provided a processor implemented method of generating a thematic summary of a document presented in machine-readable form to the processor, the document including a first multiplicity of sentences and a second multiplicity of terms, the processor implementing the method by executing instructions stored in electronic form in a memory device coupled to the processor, the processor implemented method comprising the steps

of: a) determining a value of a first number of thematic terms based upon a value of a second number representing a length of the thematic summary, the first number being less than the second number; b) selecting the first number of thematic terms from the second multiplicity of terms; c) scoring each sentence of the first multiplicity of sentences based upon occurrence of thematic terms in each sentence; and d) selecting the second number of thematic sentences from the first multiplicity of sentences based upon the score of each sentence.

**[0012]** The method of the invention automatically produces readable and semantically correct document summaries. The method requires a user, at most, to specify the length of the desired summary. Document summaries may be automatically generated without using an iterative approach.

**[0013]** A technique for automatically generating thematic summaries of machine readable documents will be described. The technique begins with identification of thematic terms within the document. Next, each sentence of the document is scored based upon the number of thematic terms contained within the sentence. Afterward, the highest scoring sentences are selected as thematic sentences. The present invention will now be described, by way of example with reference to the accompanying drawings, in which:

Figure 1 illustrates a computer system for automatically generating thematic summaries of documents, and

Figure 2 is a flow diagram of a method of generating a thematic summary of a document using the computer system of Figure 1. Figure 1 illustrates in block diagram form computer system 10 in which the present method is implemented. The present method alters the operation of computer system 10, allowing it to generate a thematic summary of any document presented in machine readable form. Briefly described, computer system 10 generates a thematic summary by identifying thematic terms within the document and then scoring each sentence of the document based upon the number of thematic terms contained within the sentence. Afterward, computer system 10 selects the highest scoring sentences as thematic sentences and presents those sentences to a user of computer system 10.

**[0014]** Prior to a more detailed discussion of the present method, consider computer system 10. Computer system 10 includes monitor 12 for visually displaying information to a computer user. Computer system 10 also outputs information to the computer user via printer 13. Computer system 10 provides the computer user multiple avenues to input data. Keyboard 14 allows the computer user to input data to computer system 10 by typing. By moving mouse 16 the computer user is able to move a pointer displayed on monitor 12. The computer user may also input information to computer system 10 by writing on electronic tablet 18 with a stylus or pen 20. Alternatively, the computer user can input data stored on a magnetic medium, such as a floppy disk, by inserting the disk into floppy disk drive 22. Optical character recognition unit (OCR unit) 24 permits the computer user to input hardcopy documents 26 into the computer system, which OCR unit 24 then converts into a coded electronic representation, typically American National Standard Code for Information Interchange (ASCII).

**[0015]** Processor 11 controls and coordinates the operations of computer system 10 to execute the commands of the computer user. Processor 11 determines and takes the appropriate action in response to each user command by executing instructions stored electronically in memory, either memory 28 or on a floppy disk within disk drive 22. Typically, operating instructions for processor 11 are stored in solid state memory 28, allowing frequent and rapid access to the instructions. Semiconductor memory devices that can be used include read only memories (ROM), random access memories (RAM), dynamic random access memories (DRAM), programmable read only memories (PROM), erasable programmable read only memories (EPROM), and electrically erasable programmable read only memories (EEPROM), such as flash memories.

**[0016]** Figure 2 illustrates in flow diagram form the instructions 40 executed by processor 11 to generate a thematic summary of a machine readable document. Instructions 40 may be stored in solid state memory 28 or on a floppy disk placed within floppy disk drive 22. Instructions 40 may be realized in any computer language, including LISP and C + +.

**[0017]** Initiating execution of instructions 40 requires selection and input of a document in electronic form. If desired, prior to initiating execution of instructions 40 the computer user may also change the length, denoted "S", of the thematic summary from the default length. The default length of the thematic summary may be set to any arbitrary number of sentences. In an embodiment intended for document browsing, the default length of the thematic summary is set to five sentences.

**[0018]** Processor 11 responds to selection of a document to be summarized by branching to step 42. During step 42 processor 11 tokenizes the selected document into words and sentences. That is to say, processor 11 analyzes the machine readable representation of the selected document and identifies sentence boundaries and the words within each sentence.

**[0019]** Tokenization of natural language text is well known and therefore will not be described in detail herein. Additionally, during tokenization, processor 11 assigns a sentence I.D. to each sentence of the document. In one embodiment, each sentence is identified by a number representing its location with respect to the start of document. Other methods of identifying the sentenc-

es may be used without affecting the present method. After tokenizing the selected document, processor 11 branches from step 42 to step 44.

**[0020]** Processor 11 examines each word token of the document during step 44 and compares the word to the terms already included in a term list. If the word token is not yet included on the list, then processor 11 adds the word to the term list and notes the sentence I.D. of the sentence in which the word occurs. On the other hand, if the word is already on the term list, processor 11 simply adds the sentence I.D. for that word token to the entry, or list, for that term. In other words, during step 44 processor 11 generates a data structure associating words of the document with the location of every occurrence of that term. Thus, for example, a term list entry of "apostasy, 7, 9, 12" indicates that the term "apostasy" occurs in sentences 7, 9, and 12 of the document.

**[0021]** Preferably, while generating the term list, processor 11 filters out stop words. As used herein, "stop words" are words that do not convey thematic meaning and occur very frequently in natural language text. Most pronouns, prepositions, determiners, and "to be" verbs are classified as stop words. Thus, for example, words such as "and, a, the, on, by, about, he, she" are stop words. Stop words within the document are identified by comparing the word tokens for the document to a list of stop words. Eliminating stop words from the term list is not necessary, but doing so reduces the total processing time required to generate a thematic summary of a document.

**[0022]** Processor 11 branches to step 46 from step 44 after completing the term list. During step 46 processor 11 analyzes the term list to determine the number of times each term occurs in the document. This is done simply by counting the number of sentence I.D.s associated with the term. That done, processor 11 branches to step 50.

**[0023]** After initiation of execution and prior to execution of step 50, during step 48, processor 11 determines the number of thematic terms to be used in selecting thematic sentences. That number, denoted "*K*", is determined based upon the length of the thematic summary; i.e., based upon S. In general, *K* should be less than *S* and greater than 1. Requiring *K* be less than S insures some commonality of theme between selected thematic sentences. Preferably, *K* is determined according to the equation:

$$K = \begin{cases} S \times c_1 & S \times c_1 > 3 \\ 3 & \text{otherwise;} \end{cases}$$

where:

$c_1$ is a constant whose value is less than 1;
S is the number of sentences in the thematic summary; and
K is the number of thematic terms.

In one embodiment, the value of $c_1$ is set equal to 0.7.

**[0024]** Armed with a value for *K* and the term counts generated during step 46, processor 11 begins the process of selecting *K* thematic terms. During step 50, processor 11 sorts the terms of the term list according to their counts; i.e., the total number of occurrences of each term in the document. Ties between two terms having the same count are broken in favor of the term including the greatest number of characters. Having generated a sorted term list and stored the list in memory, processor 11 branches from step 50 to step 52. During step 52 processor selects from the sorted term list the *K* terms with the highest counts. That done, processor 11 advances to step 54.

**[0025]** During step 54 processor 11 computes the total number of occurrences of the K thematic terms in the document. That number, denoted "*N*", is calculated by summing the counts of the *K* thematic terms. Processor 11 branches to step 56 from step 54.

**[0026]** Having selected the thematic terms and determined their counts, processor 11 is ready to begin evaluating the thematic content of the sentences of the document. During steps 56, 58, 60, and 62, processor 11 considers only those sentences that include at least one of the K thematic terms. Processor 11 does so by examining the *K* highest scoring terms of the sorted term list After selecting a term, denoted $t_s$, during step 56, processor 11 examines each sentence I.D. associated with $t_s$ during step 58. For each sentence I.D. associated with $t_s$ processor 11 increments that sentence's score. Preferably, the score for each sentence is incremented by *s*, where s is expressed by the equation:

$$s = \text{count}_{t_s}[c_2 + freq_{t_s}];$$

where:

$\text{count}_{t_s}$ is the number of occurences of $_{t_s}$ in the sentence
$c_2$ is a constant having a non-zero, positive value; and
$freq_{t_s}$ is the frequency of the selected term $t_s$.
$freq_{t_s}$ is given by the expression:
$freq_{t_s} = count_{t_s}/N$;

where:

*N* represents the total number of occurrences of thematic terms within the document.

Preferably, $c_2$ is set to a value of one.

**[0027]** Sentence scores can be tracked by generating a sentence score list during step 58. Each time processor 11 selects a sentence I.D. the sentence score list is

examined to see if it includes that sentence I.D.. If not, the sentence I.D. is added to the sentence score list and its score is increased as appropriate. On the other hand, if the sentence score list already includes the particular sentence I.D., then the score already associated with the sentence is incremented in the manner discussed previously.

**[0028]** After incrementing the scores of all sentences associated with the selected term, $t_s$, processor 11 branches from step 58 to step 60. During step 60 processor 11 determines whether all the thematic terms have been evaluated. If not, processor 11 returns to step 56 to select another thematic term as the selected term. Processor 11 branches through steps 56, 58, and 60 as described previously until all of the thematic terms have been examined. When that event occurs, processor 11 branches to step 62 from step 60.

**[0029]** During step 62 processor 11 selects as the thematic summary the *S* sentences with the highest scores. Processor 11 does this by sorting the sentence score list by score. Having selected the thematic sentences, processor 11 may present the thematic summary to the user via monitor 12 or printer 13, as well as storing the thematic summary in memory 22 or to floppy disk for later use. The sentences of the thematic summary are preferably presented in their order of occurrence within the document. While the sentences may be presented in paragraph form, presentation of each sentence individually is preferable because the sentences may not logically form a paragraph. Generation of the thematic summary complete, processor 11 branches to step 64 from step 62.

**[0030]** Thus, a method of automatically generating thematic summaries for documents has been described. The method relies upon quantitative content analysis to identify thematic words, which are used in turn to identify thematic sentences. Appendix A and Appendix B include summaries generated using this method to automatically generate thematic summaries.

## Appendix A: Summary of Shevardnadze's Resignation Speech

**[0031]** I have drawn up the text of such a speech, and I gave it to the secretariat, and the deputies can acquaint themselves with it -- what has been done is the sphere of current policy by the country's leadership, by the President and by the ministry of Foreign Affairs, and how the current conditions are shaping up for the development of the country, for the implementation of the plans for our democratization and renewal of the country, for economic development and so on.

**[0032]** Yesterday there were speeches by some comrades -- they are our veterans -- who raised the question of the need for a declaration to be adopted forbidding the President and the country's leadership from sending troops to the Persian Gulf. And these speeches yesterday, comrades, filled the cup of patience, to overflowing.

**[0033]** On about 10 occasions, both in the country and abroad, I have had to speak and explain the attitude of the Soviet Union toward this conflict.

**[0034]** In that case we would have had to strike through everything that has been done in recent years by all of us, by the whole country and by all of our people in the field of asserting the principles of the new political thinking.

**[0035]** Second, I have explained repeatedly, and Mikhail Sergeyevich spoke of this in his speech at the Supreme Soviet, that the Soviet leadership does not have any plans -- I do no know, maybe someone else has some plans, some group -- but official bodies, the Ministry of Defense -- charges are made that the Foreign Minister plans to land troops in the Persian Gulf, in the region.

**[0036]** The third issue, I said there, and I confirm it and state it publicly, that if the interests of the Soviet people are encroached upon, if just one person suffers -- wherever it may happen, in any country, not just in Iraq but in any other country -- yes, the Soviet Government, the Soviet side will stand up for the interests of its citizens.

**[0037]** I say that, all the same, this is not a random event. Excuse me, I am now going to recall the session of the supreme soviet. On comrade Lukyanov's initiative, literally just before the start of a meeting, a serious matter was included on the agenda about the treaties with the german democratic republic.

**[0038]** I cannot reconcile myself with what is happening in my country and to the trials which await our people.

## Appendix B: Summary of "Research that Reinvents the Corporation" by John Seely Brown

**[0039]** As companies try to keep pace with rapid changes in technology and cope with increasingly unstable business environments, the research department has to do more than simply innovate new products.

**[0040]** Over the next decade, PARC researchers were responsible for some of the basic innovations of the personal-computer revolution-only to see other companies commercialize these innovations more quickly than Xerox.

**[0041]** One popular answer to these questions is to shift the focus of the research department away from radical breakthroughs toward incremental innovation, away from basic research toward applied research.

**[0042]** Our emphasis on pioneering research led us to redefine what we mean by technology, by innovation, and indeed by research itself.

**[0043]** Such activities are essential for companies to exploit successfully the next great breakthrough in information technology: "ubiquitous computing," or the incorporation of information technology in a broad range of everyday objects.

**[0044]** When corporate research begins to focus on a

company's practice as well as its products, another principle quickly becomes clear: innovation isn't the privileged activity of the research department. At PARC, we are studying this process of local innovation with employees on the front lines of Xerox's business and developing technologies to harvest its lessons for the company as a whole.

**[0045]** The result: important contributions to Xerox's core products but also a distinctive approach to innovations with implications far beyond our company.

## Claims

1. A processor implemented method of generating a thematic summary of a document (26) presented in machine-readable form to the processor (11), the document (26) including a first multiplicity of sentences and a second multiplicity of terms, the processor (11) implementing the method by executing instructions stored in electronic form in a memory device (28) coupled to the processor (11), the processor implemented method comprising the steps of:

   a) determining (48) a value of a first number of thematic terms based upon a value of a second number representing a length of the thematic summary, the first number being less than the second number;

   b) selecting (52) the first number of thematic terms from the second multiplicity of terms;

   c) scoring (58) each sentence of the first multiplicity of sentences based upon occurrence (54) of thematic terms in each sentence; and

   d) selecting (62) the second number of thematic sentences from the first multiplicity of sentences based upon the score of each sentence.

2. The processor implemented method of claim 1 wherein step b) comprises:

   i) determining (46) a number of times each term of the second multiplicity of terms occurs in the document, and

   ii) selecting (52) the first number of thematic terms from the second multiplicity of terms based upon the number of times (50) each term occurs in the document.

3. The processor implemented method of claim 1 or claim 2 further comprising the step of:

   e) presenting the thematic sentences to a user of the processor (11) in an order in which the thematic sentences occur in the document.

4. The processor implemented method of any one of claims 1 to 3 wherein step c) comprises incrementing (58) the score of each sentence for each thematic term occurring in the sentence by an amount related to the frequency of occurrence of the thematic term within the document.

5. The processor implemented method of any one of claims 1 to 4 comprising, prior to step a), the step of receiving the value of the second number from an input device (14) coupled to the processor (11).

6. The processor implemented method of any one of claims 1 to 5 wherein the first number is at least three.

## Patentansprüche

1. Prozessorimplementiertes Verfahren zum Erzeugen einer thematischen Zusammenfassung eines Dokuments (26), das dem Prozessor (11) in maschinenlesbarer Form präsentiert wird, wobei das Dokument (26) eine erste Vielzahl von Sätzen und eine zweite Vielzahl von Termen umfasst und der Prozessor (11) das Verfahren durch Ausführen von Instruktionen implementiert, die in elektronischer Form in einer Speichereinrichtung (28), die mit dem Prozessor (11) gekoppelt ist, gespeichert sind, wobei das prozessorimplementierte Verfahren die folgenden Schritte aufweist:

   a) Bestimmen (48) eines Wertes einer ersten Anzahl von thematischen Termen beruhend auf einem Wert einer zweiten Anzahl, die eine Länge der thematischen Zusammenfassung repräsentiert, wobei die erste Anzahl kleiner als die zweite Anzahl ist;

   b) Auswählen (52) der ersten Anzahl von thematischen Termen von der zweiten Vielzahl von Termen;

   c) Bewerten (58) jedes Satzes der ersten Vielzahl von Sätzen beruhend auf eines Auftretens (54) von thematischen Termen in jedem Satz; und

   d) Auswählen (62) der zweiten Anzahl von thematischen Sätzen von der ersten Vielzahl von Sätzen beruhend auf der Bewertung jedes Satzes.

2. Prozessorimplementiertes Verfahren nach Anspruch 1, wobei Schritt b) aufweist:

i) Bestimmen (46) einer Häufigkeit, mit der jeder Term der zweiten Vielzahl von Termen in dem Dokument auftritt, und

ii) Auswählen (52) der ersten Anzahl von thematischen Termen von der zweiten Vielzahl von Termen beruhend auf der Häufigkeit (50), mit der jeder Term in dem Dokument auftritt.

3. Prozessorimplementiertes Verfahren nach Anspruch 1 oder 2, des Weiteren aufweisend den Schritt:

e) Präsentieren der thematischen Sätze zu einem Benutzer des Prozessors (11) in einer Reihenfolge, in der die thematischen Sätze in dem Dokument auftreten.

4. Prozessorimplementiertes Verfahren nach einem der Ansprüche 1 bis 3, wobei Schritt c) ein Inkrementieren (58) der Bewertung von jedem Satz umfasst, für jeden thematischen Term, der in dem Satz auftritt, um einen Betrag, der sich auf die Häufigkeit des Auftretens des thematischen Terms innerhalb des Dokuments bezieht.

5. Prozessorimplementiertes Verfahren nach einem der Ansprüche 1 bis 4, umfassend, vor Schritt a), den Schritt des Empfangens des Werts der zweiten Anzahl von einer Eingabeeinrichtung (14), die mit dem Prozessor (11) gekoppelt ist.

6. Prozessorimplementiertes Verfahren nach einem der Ansprüche 1 bis 5, wobei die erste Anzahl mindestens 3 ist.

## Revendications

1. Procédé mis en oeuvre par un processeur de création d'un résumé thématique d'un document (26) présenté sous forme pouvant être lu par une machine au processeur (11), le document (26) comprenant une première variété de phrases et une seconde variété de termes, le processeur (11) mettant en oeuvre le procédé en exécutant les instructions stockées sous forme électronique dans un dispositif à mémoire (28) accouplé au processeur (11), le procédé mis en oeuvre par un processeur comprenant les étapes :

a) de détermination (48) d'une valeur d'un premier nombre de termes thématiques sur la base d'une valeur d'un second nombre représentant une longueur du résumé thématique, le premier nombre étant inférieur au second nombre ;

b) de sélection (52) du premier nombre de termes thématiques à partir de la seconde variété de termes ;

c) d'évaluation (58) de chaque phrase de la première variété de phrases sur la base de l'apparition (54) de termes thématiques dans chaque phrase ; et

d) de sélection (62) du second nombre de phrases thématiques à partir de la première variété de phrases sur la base de l'évaluation de chaque phrase.

2. Le procédé mis en oeuvre par un processeur selon la revendication 1, dans lequel l'étape b) comprend :

i) la détermination (46) d'un nombre de fois que chaque terme de la seconde variété de termes apparaît dans le document, et

ii) la sélection (52) du premier nombre de zermes thématiques à partir de la seconde variété de termes sur la base du nombre de fois (50) que chaque terme apparaît dans le document.

3. Le procédé mis en oeuvre par un processeur selon la revendication 1 ou la revendication 2 comprenant en outre l'étape de :

e) présentation des phrases thématiques à un utilisateur du processeur (11) dans un ordre dans lequel les phrases thématiques apparaissent dans le document.

4. Le procédé mis en oeuvre par un processeur selon l'une quelconque des revendications 1 à 3, dans lequel l'étape c) comprend l'incrémentation (58) de l'évaluation de chaque phrase pour chaque terme thématique apparaissant dans la phrase d'une quantité relative à la fréquence d'apparition du terme thématique dans le document.

5. Le procédé mis en oeuvre par un processeur selon l'une quelconque des revendications 1 à 4, comprenant, avant l'étape a), l'étape de réception de la valeur du second nombre à partir d'un dispositif d'entrée (14) accouplé au processeur (11).

6. Le procédé mis en oeuvre par un processeur selon l'une quelconque des revendications 1 à 5, dans lequel le premier nombre est au moins trois.

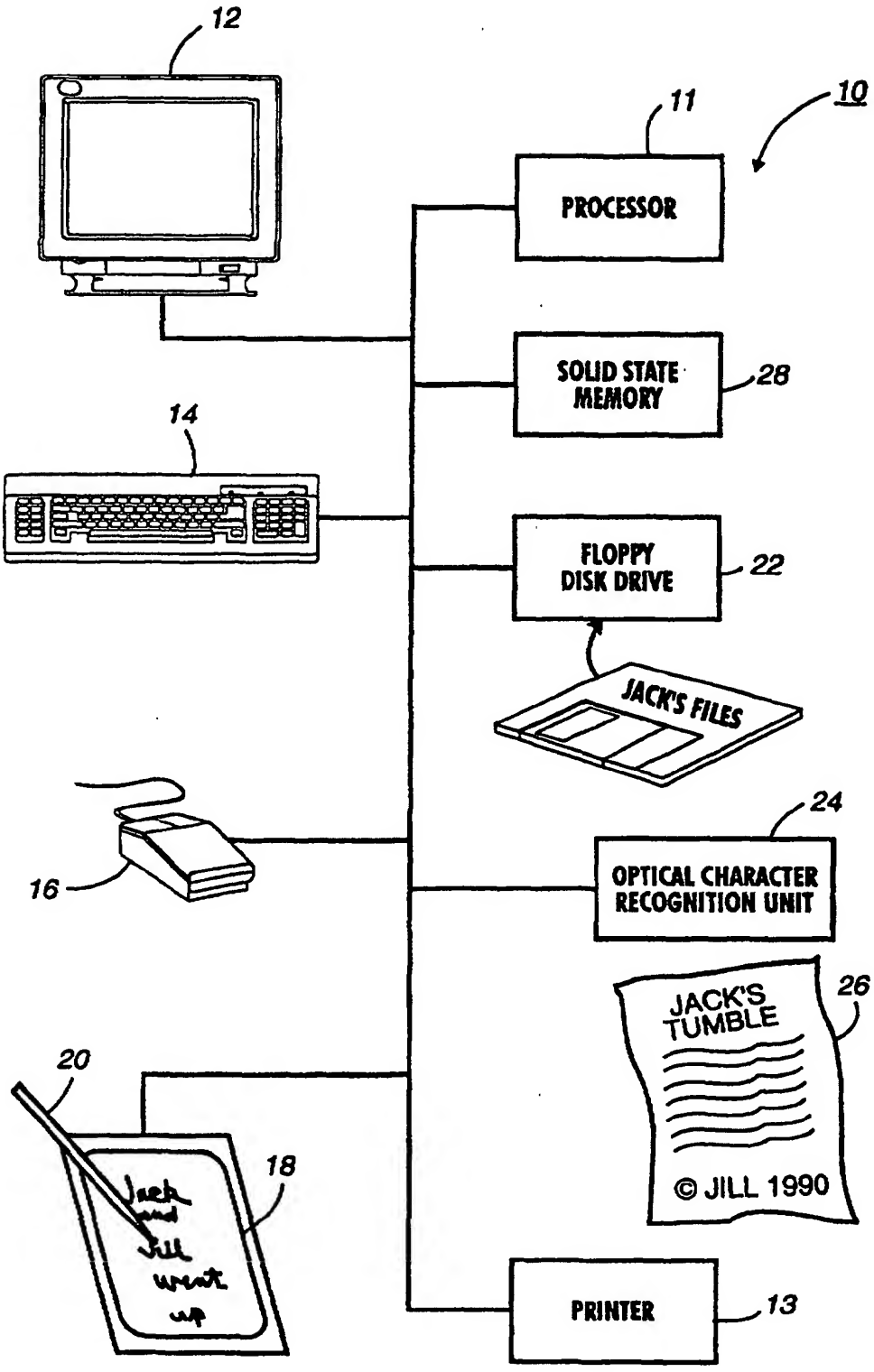
12
11
10
PROCESSOR
SOLID STATE MEMORY
28
14
FLOPPY DISK DRIVE
22
JACK'S FILES
16
24
OPTICAL CHARACTER RECOGNITION UNIT
JACK'S TUMBLE
26
© JILL 1990
20
18
PRINTER
13

FIG. 1

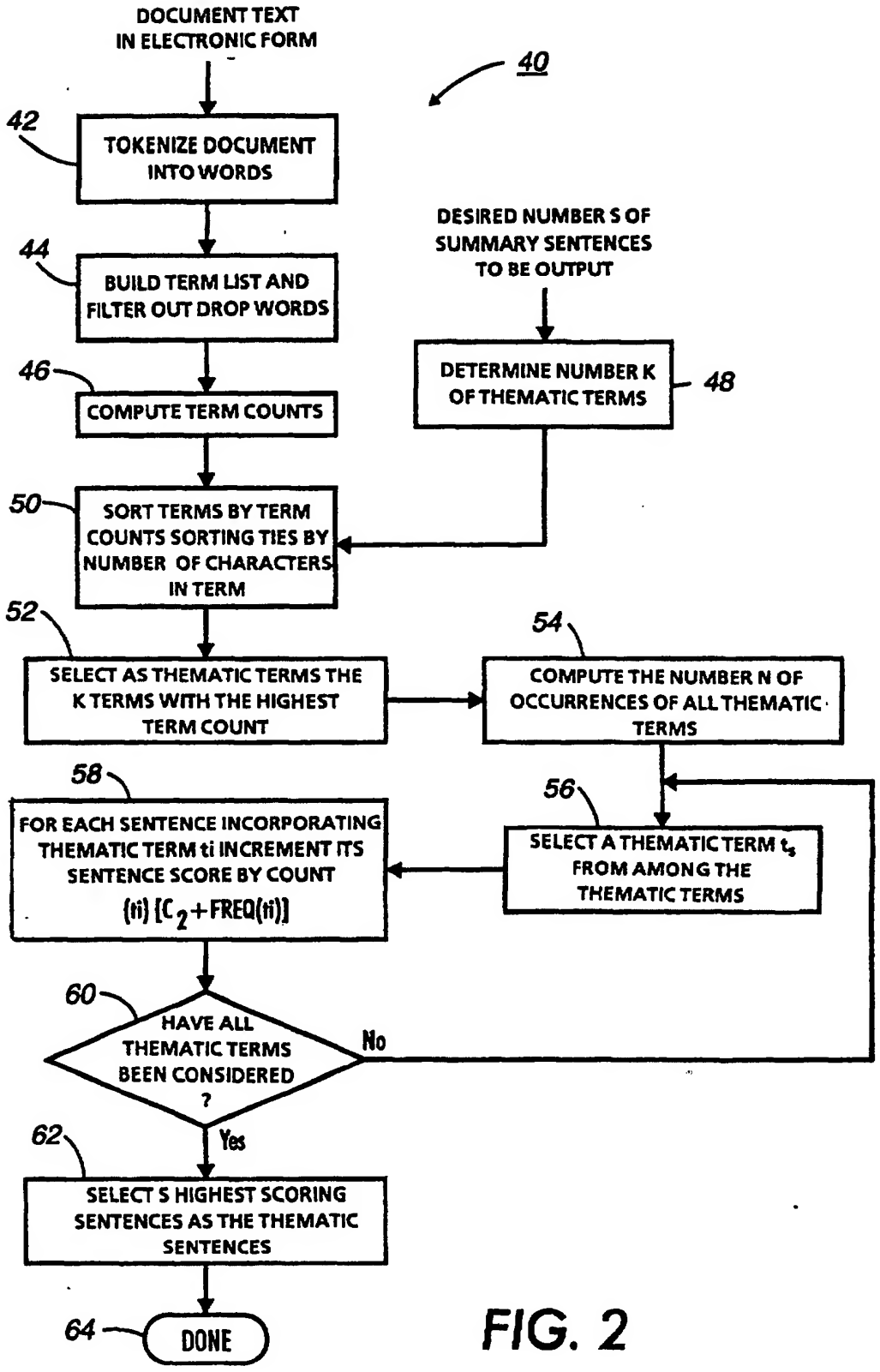
DOCUMENT TEXT IN ELECTRONIC FORM
40
42
TOKENIZE DOCUMENT INTO WORDS
DESIRED NUMBER S OF SUMMARY SENTENCES TO BE OUTPUT
44
BUILD TERM LIST AND FILTER OUT DROP WORDS
46
COMPUTE TERM COUNTS
DETERMINE NUMBER K OF THEMATIC TERMS
48
50
SORT TERMS BY TERM COUNTS SORTING TIES BY NUMBER OF CHARACTERS IN TERM
52
SELECT AS THEMATIC TERMS THE K TERMS WITH THE HIGHEST TERM COUNT
54
COMPUTE THE NUMBER N OF OCCURRENCES OF ALL THEMATIC TERMS
58
FOR EACH SENTENCE INCORPORATING THEMATIC TERM ti INCREMENT ITS SENTENCE SCORE BY COUNT
(ti) [C2+FREQ(ti)]
56
SELECT A THEMATIC TERM ts FROM AMONG THE THEMATIC TERMS
60
HAVE ALL THEMATIC TERMS BEEN CONSIDERED ?
No
Yes
62
SELECT S HIGHEST SCORING SENTENCES AS THE THEMATIC SENTENCES
64
DONE

FIG. 2